计算机汉字信息处理

姚天顺
王宝库 编著
颜秀英

辽宁科学技术出版社

计算机汉字信息处理

姚天顺　王宝库　颜秀英 编著

辽宁科学技术出版社
一九八六年·沈阳

内 容 提 要

本书阐述和总结了当前国内外计算机汉字信息处理的主要研究领域和成果，包括作者自己的工作。书中介绍了汉字编码、代码转换、字形存贮和信息压缩、汉字支撑软件、中文数据库、汉语语言理解和汉字自动识别等方面的内容和方法。本书既是一本入门书，又为读者提供了许多汉字信息处理的具体方法，且易于移植到自己的计算机上去。当然某些内容还有待于读者进一步开发研究。

本书可供从事中文信息研究、计算机科学、管理信息工程方面科技人员阅读，也可作为高等学校有关专业大学生或研究生40学时的教材或教学参考书。

计算机汉字信息处理

Jisuanji Hanzi Xinxi Chuli

姚天顺 王宝库 颜秀英 编著

辽宁科学技术出版社出版 （沈阳市南京街6段1里2号）
辽宁省新华书店发行 朝阳新华印刷厂印刷

开本：787×1092 1/32 印张：6 3/4 字数：150,000
1986年11月第1版 1986年11月第1次印刷

责任编辑：马凤兰 封面设计：秀 中

印数：1—5,800
统一书号：15288·176 定价：1.40元

序　言

电子计算机的问世，至今仅有三十多年的历史，它以惊人的速度发展起来，并有力地推动着科学技术、经济和军事科学的迅猛发展。这项二十世纪的重大发明，已成为当今世界衡量一个国家现代化水平的重要标志。它的应用技术涉及到经济管理、过程控制、军事技术、农业科学和文化教育等各个领域。

计算机传递信息，习惯上采用西文、数字和其他一些特殊符号，这对于一直沿用汉字进行日常交往的我国来说，是不方便的，难于被大多数人接受和掌握。因而解决汉字信息的计算机处理，对于在我国推广应用计算机技术是非常重要的。

汉字是当今世界各国正在使用的文字中最古老的文字，已有三千多年的历史。它在我国及亚洲一些国家文化史上建立了不可磨灭的功勋，是人类的宝贵财产。

在利用计算机进行信息处理时，汉字，作为一种交往信息的载体，有许多优点，例如存贮容量小，同样一篇文章，一般情况中文可以比英文省1/4～1/3左右。汉字是单音节的文字，语音识别可能比多音节的西文方便。又如中文的每个汉字，是一个二维图形，大多数汉字本身都具有一定的语义和语音信息。这对将来进入第五代计算机，处理人—机系统和自然语言理解等方面将会带来便利条件。

当然，汉字这样古老的文字，对于应用计算机进行文字处理方面也存在一些不利之处。例如中文打字的速度不及英文打字速度的十分之一，是世界上效率最低的。在国际上，以汉语作为第二语言的人数，不到万分之一，是世界上十大语言中通用率最低的。外国人学中文是很困难的，他们从汉语拼音开始，一个字一个字地记，文法也比较困难，可以说很少有几个外国学者能够得心应手地掌握汉字。这给不同国家间的信息交换带来较多的不便。

七十年代中叶，世界上有10万种定期的科技学报，每年在学报上刊登150多万篇科技文献，2万6千多种定期性科技索引，每年出版25万余种科技书籍用60种文字印刷，其中英文占50.5%，而中文却少于0.5%。我国七十年代平均每年印刷出版刊物约有30亿字，但译成英、俄、法、德等外国文字并出版的，不到这个数量的1%。其原因除了我们的出版事业落后之外，文字所带来的困难也是重要因素之一。

特别是以西文为基础而发明的电子计算机的出现，对于中文的处理有很多极不协调的地方，需要进行系统的改造。例如，按照计算机的输入要求需要设计一套汉字编码，输出需要标准字库；此外又如，汉字的属性库和字模库、中文文本编辑、中文数据库、中文办公室系统等等一系列与汉字信息计算机处理直接有关的基础工作和应用开发都需进一步研究。

由于中文计算机的开发，古典的传统方法已满足不了日益变化的要求，也直接影响着中文语言学的发展。而语言分析和结构语言学、形式语言学、算法语言学和数理语言学等等语言学的现代数学方法，把语言学和计算机科学结合起来，构造了一个全新的领域，形成一个边缘科学。促使我们

实现汉字信息处理的现代化。

从六十年代开始，使用汉字的国家之一的日本，在汉字的机械化和自动化方面做了一些工作，他们首先把电子计算机引进了汉字信息处理领域。从新闻出版业开始，建立汉字终端，确定了信息交换用的汉字编码，并扩大应用，逐步推进到汉字情报检索、商业系统、字形识别、机器翻译等很多方面。在当时，取得世界的领先地位。以日本的《经济新闻》为例，这是一种每天出一百多版的日报。报社内安装两台大型电子计算机系统作为中央控制中心，遍设大量终端。在新闻稿的收集、原稿的修改加工、编辑处理、版面设计、印刷校样、制版、报纸印刷、印出用户标签、分打邮包，直至最后由传送带自动送上等候在出口端的各自邮车等，全都由电子计算机加以控制，全过程仅在三小时内完成，实现了高度自动化。

除日本外，美国、英国、法国、联邦德国、澳大利亚、新西兰、新加坡、加拿大、苏联、南朝鲜各国、香港等地以及台湾省等许多学者从六十年代开始，就致力于汉字信息处理技术的研究。到了七十年代，已有多种产品问世。日本和南朝鲜的汉字信息处理系统已进入实用化阶段。

我国从六十年代开始，也有不少学者致力于汉字信息处理的工作，提出了不少有价值的编码方案。但是，较大规模的研究，还是在七十年代后期开始的，汉字信息处理系统逐步在我国的一些地区和部门开始试制和试用。中国仪器仪表学会于1980年2月在苏州首先成立了汉字信息处理系统研究会。1981年6月又成立了中国中文信息研究会。标志着中文信息研究的基础更加广泛，内容更加丰富，进入协同作战的新阶段。

关于中文信息的研究，我们在设备方面虽然还落后于某些国家，但是在编码等方面已居世界领先地位。汉字系统研究的范围是很广的，从编码开始，解决汉字输入输出，创造汉字智能终端，中文办公室系统，汉字的印刷体和手写体识别，汉语语音识别，非过程智能语言，中文自然语言处理，知识理解以及国民经济和军事技术等各方面的信息管理系统等等。我国目前的情况，已具有开发计算机的研究和应用的必备条件，中文汉字信息作为一个系统的学术领域，正受到广泛的重视。

本书共五章，第一章回顾了编码的历史，重点介绍几种典型的编码方法，以及实现各种编码方案的输入码机内转换和国标码转换方法，并讨论了汉字信息处理的输入问题。第二章是输出问题，即汉字库的建立和字形的信息压缩，包括可逆压缩和不可逆压缩，以实现存储量小输出字形美观的目的。第三章是在解决了输入输出的基础上，研究中文文本编辑，中文高级语言和汉字数据库。目的在于构造一个汉字信息处理的环境，提供汉字信息处理的支撑软件，以便更好地开发应用系统。第四章是汉语自然语言理解问题。这是人工智能的核心课题，也是一个非常有兴趣的课题。我们的重点在于介绍根据Winograd的ATN网络和Schank 的概念从属理论构造起来的汉语理解模型。这很不成熟，主要提供讨论。最后一章简要地介绍几种汉字自动识别问题。它也是一个重大的课题，目前虽有所进展，但有待于我们开发和继续研究。

本书初稿作为“汉字信息处理”这门课程的讲义，曾两度在东北工学院为大学生和研究生讲授，有关老师和学生对此进行了评议，特别是研究生周鹤人、滕永林、邢国良、李

长山等做了不少工作，提出了有益的建议，作者又作了修改，在此真诚地表示感谢。尽管如此，由于研究工作和水平所限，错误和问题仍在所难免，请读者批评指正。

作　者

一九八五年五月

目　录

第一章　汉字的编码

汉字信息的处理和编码是实现中文信息计算机处理的关键问题之一，解决了这个问题，汉字才能高效率地、准确地输入给计算机。

汉字编码属于边缘科学，涉及多种学科。这是一个研究了十多年而仍未彻底解决的问题，必须认真地对待。为此，我们先作一些历史的回顾。

第一节　历史的回顾

从殷商时代（公元前十四世纪）的甲骨文算起，汉字已有三千五百年的历史，仅次于古埃及的圣书字和古巴比伦的丁头字。在世界各国现行文字里，汉字是历史最悠久的文字之一，它在漫长的历史进程中不断地变化和发展，如汉字的总字数多次增加，字形几度简化，读音逐渐更改等。

汉字的字种数目究竟有多少？根据有关资料记载，殷代的甲骨文和殷周金文只有两千字左右。东汉末年（公元 121 年）许慎的《说文解字》收入9,353字。清朝的《康熙字典》（1715年）增加到 42,174 个字。1915年出版的《中华大字典》有44,908个字。不过这些字不都是经常使用的，我们日常使用的大概有五、六千字就足够了。

汉字的编码历史，可以分为三个发展阶段。

一、编纂字典、词典的萌芽阶段

在这个阶段，人们主要是为了汉字的查询求解，编纂字典词典，将汉字编码排序，以利使用。如春秋时期的《尔雅》，东汉的《说文解字》，清朝的《康熙字典》，近代的各种“字典”、“词源”、“辞海”等。就编码的方法而言，有传统的部首笔画编码，王云五的四角笔形编码，丁西林的笔形查字法，周辨明的半周钥笔索引法，以及拼音排序编码等。这个阶段编码的特点是用偏旁部首对文字进行分类，用笔画多少对文字进行排序，并发明了切韵双拼和四声来标注读音。

二、近代电报码阶段

由于通商贸易发展的需要，1880 年满清政府创办电报局，在天津到塘沽和天津到上海的电路上试验电报通讯。当时雇佣丹麦人制定了汉字的电报编码，后来逐步发展成为现在的标准电码本。最初，电报码用的是四位数字等长码，以后又有三位英文字母等长码、气象电报码以及各种密报码等。这个阶段编码的特点是把汉字编成代码，一个汉字用一组数码或拉丁字母来表示。

三、现代计算机输入编码阶段

利用计算机处理汉字，以英数符代码为输入的编码方案，近年来有很大的发展，现约有几百种，归纳起来有三种类型：①拼音码，类似于汉语拼音，利用英文字母和数字作编码。②音形码，也可以称为声韵与部、形、文、频度结合码。例如在双拼标调的基础上再加部首码、形码、字义区别

符或字顺顺序符来制定编码方案，或者把拼形的部件（字根、字元）用部件的读音代表符号来作编码。③纯形码（或拼形码），整字输入，或者从分析汉字的形态结构中由汉字的部件组拼。也有人采用四角码、三角码，即用英文字母或数字作代表符来编码。

汉字用键盘输入，按键盘上键位的多少可划分为大、中、小三种。一般说：大键盘整字输入，中键盘属于字元组拼输入，小键盘按拼音或笔画字元输入。

汉字编码是中文信息处理的最基本的问题，正受到国内外的广泛注意。日本、南朝鲜、新加坡、美国、加拿大、英国、联邦德国、澳大利亚、台湾省和香港地区等都提出了一些有价值的方案。下面，我们对几个典型的方案作一些较详细的介绍。

第二节　汉字编码的典型方法

一、汉语拼音方案的编码方法

这是一种比较典型的拼音编码方案。字母表、声母表、韵母表和声调符号均沿用汉字拼音方案。组拼时不必使用隔音符号，因为汉字编码是按字区分的，不会出现音节界限混淆。拼音方案的最大问题是同音字太多。为了区别同音字，选取20字偏旁字母，如表1·1所示。

表1·1　偏旁字母表

偏旁字母	人	才	口	阝(卩)	忄	钅(金)	木	氵
名　称	rén 人	shǒu 手	kǒu 口	ěr 耳	xīn 心	jīn 金	mù 木	shuǐ 水

偏旁字母	火(灬)	土	日	月(曰)	⺾	虫
名 称	huǒ 火	tǔ 土	rì 日	yuè 月	cǎo 草	chóng 虫
偏旁字母	纟(幺)	又	讠	女(夕)	山	个(竹)
名 称	sī 丝	yòu 又	yán 言	nǚ 女	shān 山	zhú 竹

偏旁用作区分符号，置于拼音字之前，不作其他用。例如“人REN”表示“仁”。下面我们列举一些例子加以说明：

汉 字	拼 音 码
肯	kěn
啃	口 kěn
垦	𠂆 kěn
垦	土 kěn

汉 字	拼 音 码
半	bàn
拌	扌 bàn
样	木 bàn
伴	人 bàn
绊	纟 bàn
扮	又 bàn
办	土 bàn
瓣	⺾ bàn

按《新华词典》考虑，偏旁部首总共有一百八十多个，但是该方案仅取二十个。所以从上述例子可以看到：在同音字较多的情况下，有些字的组拼还不能根据人们的习惯来

取，必须给出其他一些特殊的规定，将所有的偏旁都归并压缩在20个以内。但实际作起来，是有一定困难的。

二、“见字识码”的编码方法

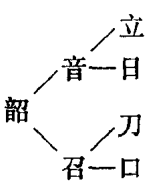
这种编码方法是由中国科学院学部委员支秉彝博士提出来的，受到国内外有关人士的普遍重视。

在正式讨论这个编码方案之前，先讨论一下汉字的结构和字元。

汉字的结构层次分为三级：整字、字元、笔画。其中整字和笔画的概念是很清楚的，而字元（亦称“字根”或“部件”）是一种具有固定形体，明确称读和一定含义的构字基本单位。

字元，根据形体特点分为单元（如“木、目”）和复元（如“旻、穴”），根据音义特点分为成字字元（如“口、日”）和不成字字元（如“亻、讠”）。

一个汉字分解成字元，应从形体上分解，兼顾音义。因为汉字是形音义的统一体，一般可以采用“层次二分法”，分解到单元为止。如：



从分解过程的图形结构上考虑，其基本拓扑图形可以定为四种：单体型、左右型、上下型和外内型。韶字是左右型，音字为上下型，日字是单体型，国字就是外内型。

该方案就是建立在字音和字形的双重关系上，把字拆分

成字元，力图符合人们的习惯，做到“见字识码”。

1．编码的基本原则

方案从人们的习惯出发，把字拆分成字元，每个字规定由四个字元组成，字元表如表1·2所示。

表1·2　字元、关系字和拼音汉字对应表

字元						关系字	拼音汉字
一画							
丶						点	Dian
丨						直	Zhi
亅	𠂉	𠃌				勾	Gou
𠂇						提	Ti
丿						撇	Pie
乀						捺	Na
二画							
冖						幂	Mi
冫	氵					冷	Leng
匕	匚	卜				匕	Bi

工						工	Gong
年	年					年	Nian
直	直	直	直	直		直	Zhi
亻	人	人				人	Ren
冂	冂	月	同	同		同	Tong
丷	丶					八	Ba
辶	辶					走	Zou
阝	阝					耳	Er
刀	刀					刀	Dao
六						六	Liu
左	左					左	Zuo
匚	匚					巨	Ju
厂	厂	厂				厂	Chang
凶	凶					凶	Xiong